Universidad de Los Andes

Big Data and Machine Learning for Applied Economics

MECA 4107

Junio 26,2022

Bogotá D.C, Colombia

**PROBLEM SET 1: PREDICTING INCOME**

Maria Valeria Gaona Guevara

Código: 202214418

Correo: [mv.gaona@uniandes.edu.co](mailto:mv.gaona@uniandes.edu.co)

Andrea Margarita Beleño Hernández

Código: 200620739

Correo: [a.beleno@uniandes.edu.co](mailto:a.beleno@uniandes.edu.co)

**Resumen:** El siguiente documento presenta XXXXXXXX. El repositorio GitHub del siguiente documento se encuentra en el siguiente link:

1. **Información General**
   1. ***Adquisición de datos.***

**El objetivo principal es construir un modelo predictivo de la renta individual:**

***Ingreso= f(*X) *+ u***

* + 1. **Extraer los datos que se encuentran disponibles en el siguiente sitio web https://ignaciomsarmiento.github.io/GEIH2018\_sample/.**

Para realizar la importación de datos es importante conocer qué tipo de página web es. La página web que contiene la base de datos es dinámica, haciendo la extracción de datos sea más compleja. Los datos fueron extraídos de cada uno de los 10 enlaces presentados en el sitio web. El código utilizado se encuentra en el archivo: “GEIH-Archivo de trabajo oficial.r” y la base de datos en la carpeta: “Elementos guardados” con el nombre: “Datos\_GEIH.rds”, luego de hacer una inspección previa del código html de la página para determinar la fuente de los datos. En las Figuras 1 y 2, se muestra que al realizar la inspección, se encuentra que la página no cuenta con la tabla de datos directa, por lo cual se procedió a revisar el enlace de la página donde residen los datos, los cuales se encuentran contenidos en el Script.



*Figura 1. Código fuente de la página web – página fuente de datos*



*Figura 2. Código fuente de la página web- función para incluir la tabla*

Dentro de la realización del script en R, es necesario descargar un paquete llamado *rvest,* el cual se utiliza como herramienta para hacer el Web Scraping, es decir, importar los datos desde una página web.

1. Dado el tamaño de cada base datos, se procedió con la descarga desde la página web específica y cada base de datos se convirtió en matriz:

*Base1<read\_html("https://ignaciomsarmiento.github.io/GEIH2018\_sample/pages/geih\_page\_1.html")%>% html\_table()*

*Base1 <- data.frame(Base1)*

*Base1<read\_html("https://ignaciomsarmiento.github.io/GEIH2018\_sample/pages/geih\_page\_2.html")%>% html\_table()*

*Base2 <- data.frame(Base2)*, hasta terminar de descargar las 10 bases de datos.

1. Posteriormente, se genera una fusión entre las 10 bases datos para que se conforme la GEIH completa

*DatosGEIH<- rbind(Base1, Base2, Base3, Base4, Base5, Base6, Base7, Base8, Base9, Base10)*. Con este código se obtuvo la base de datos completa con 178 variables y 32177 observaciones.

* + 1. **¿Existen restricciones para acceder o extraer estos datos?**

Para acceder a los datos es necesario esperar un momento, ya que los datos no se cargan automáticamente, debido a que los datos no se encuentran en ese mismo enlace web, sino que los extrae de otra página web donde realmente reposan los datos. De acuerdo con lo anterior, sí existen restricciones para extraer directamente los datos debido al inconveniente mencionado anteriormente; cuando se corre el código de extracción de datos en R, no detecta la tabla en html que se visualiza en la página web, porque esta se importa a través de la función “*includehtml()*”, que se encuentra en el script. Sin embargo, al obtener el enlace de página web en donde reposan los datos en html, no existen restricciones para descargar o realizar “scraping” a los datos.

* + 1. **Usando pseudocódigo, describa su proceso de adquisición de datos**

1. Ingresar al enlace: https://ignaciomsarmiento.github.io/GEIH2018 muestra/.
2. Ingresar al enlace: Data chunk 1 (<https://ignaciomsarmiento.github.io/GEIH2018_sample/page1.html>)
3. Realizar la inspección del código fuente de la página para determinar si la tabla existe.
4. Al verificar que la tabla no se encuentra, verificar la fuente de la tabla de datos en html.
5. Obtener el enlace a la tabla donde se encuentra la tabla de datos en html.
6. Validar que para los demás datasets el enlace a la página web.
7. Elaborar el código en R para realizar la extracción precisa de los datos
   1. Instalar los paquetes necesarios
   2. Realizar el vector de los enlaces de las páginas web donde se extraerán los datos
   3. Realizar el scrape de los datos usando la función de “rvest” [readhtml () y html\_table()]
   4. Unir los datos en una base de datos de extensión “.rds”
   5. **Limpieza de datos**

**1.2.1 El set de datos incluye múltiples variables que pueden ayudar a explicar el ingreso individual. Guiado por su intuición y conocimiento en economía, escoja las variables más relevantes y realice un análisis descriptivo de estas variables. Por ejemplo, puede incluir variables que midan la educación y la experiencia, dadas las implicaciones del modelo de acumulación de capital humano (Becker, 1962, 1964; y Mincer (1962, 1975)**

Para conocer el ingreso total es fundamental contar con las siguientes variables:

1. Pet = Población en edad de trabajar
2. Edad
3. Educ = Hace referencia a la educación con la que se cuenta. En la base de datos se trabajará la variable p6210, en donde se evidencia el nivel de educativo más alto alcanzado
4. Ocu= Ocupación
5. Sex = Género.
6. Exp= Experiencia. En la base de datos se trabajará la variable p6426, en donde se dividirá este dato entre 12 para tener los datos en formato años.

A continuación, se describen las variables para tener en cuenta de la base de datos:

*Edad y población en edad de trabajar. (Edad y Pet)*

La edad de un individuo tiende a representar sus necesidades, oficios, intereses y preferencias. Por lo tanto, conocer la edad de los individuos nos permite generar un filtro para observar cuál es la población objetivo para cada investigación y planteamiento que se desee presentar. En este modelo de ingresos, los menores de edad, por ejemplo, no representan información representativa, ya que cuentan usualmente con un jefe de hogar quien es el que percibe sus ingresos para manutención y demás necesidades. Por consiguiente, sus preferencias, oficios e intereses no serán analizadas en este espacio.

De acuerdo con lo anterior, es fundamental contar con una segmentación por edades, ya que eso permite contar con un panorama más claro para proceder con el análisis. La población en edad de trabajar representa aquellos individuos que pueden generar ingresos por concepto de trabajo y ser jefes de los hogares, haciendo que, esta variable sea necesaria para contar con un modelo objetivo y claro sobre cuál será la población a describir

*Educación (Educ)*

La educación representa cuán capacitado y certificado está el individuo. La educación le permite al individuo poder contar con mejor salario, ya que se asume que el individuo entre más educación posea, es más competente y con ello, tiende a ser más productivo. Por lo tanto, contar con la educación en el modelo, permite analizar cuán importante es la educación para saber cuántos ingresos puede a llegar a obtener si aumenta uno o más años de estudio.

*Ocupación (Ocu)*

La ocupación permite filtrar a aquellos individuos que se encuentran en edad de trabajar y en este momento se encuentran con trabajo, u ocupados. Esto permite contar con un espectro más claro en el modelo, ya que permite analizar cómo sus ingresos dependen si están ocupados o no.

*Género (Sex)*

El género es fundamental en el análisis de los ingresos de los individuos, ya que en el contexto Colombiano, por ejemplo, existe una brecha entre hombres y mujeres en el momento de obtener trabajo y ganar un salario determinado, por lo tanto, ser hombre o mujer sí tiene influencia en la cantidad de ingresos que se perciben. Es por eso que en el modelo tiene que estar presente esta variable, ya que ayudará a conocer el impacto en el salario dependiendo del género que tenga dicho individuo.

*Experiencia (Exp)*

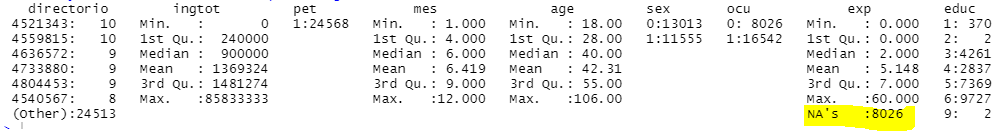
La experiencia permite conocer cuánto tiempo ha durado una persona trabajando, en este caso, la base de datos nos presenta los datos del tiempo que lleva trabajando la persona en la empresa actual. Esta variable es muy importante, ya que aporta al conocimiento la influencia que tiene la experiencia al ingreso de una persona, ya que entre más tiempo lleve trabajando, puede ser más productivo porque cuenta con más conocimiento y habilidades para realizar sus actividades.

*Directorio (Directorio)*

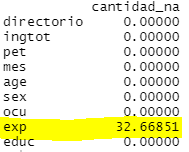
El directorio nos permite establecer si las personas a las que se les realizó la encuesta viven en un mismo hogar. Esto será relevante en la medida que se puede establecer si las personas que viven en el mismo hogar pueden ser susceptibles de tener mayores ingresos.

***1.2.2* Observe que hay muchas observaciones con datos faltantes. Dejo en sus manos el encontrar la manera de tratar con estos datos. En su explicación, describa los pasos que usted realizó para limpiar los datos y justifique su decisión.**

Para las variables mencionadas anteriormente, existen datos faltantes para la variable “exp”, como se presenta en la tabla a continuación:



Así mismo, se analiza el porcentaje de NA incluido en la variable “exp”, en donde se observa que el 32% de la base de datos contiene NA:



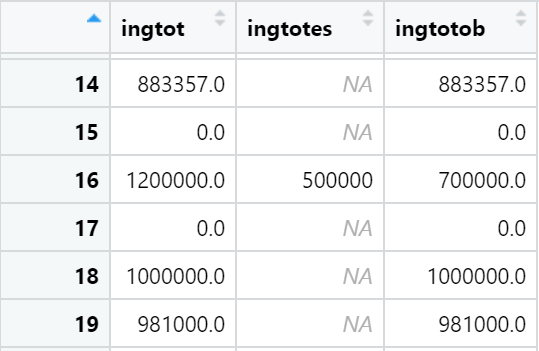
Para la variable “exp”, si bien tiene una gran cantidad de NA, y si se fuese a hacer una regresión lineal, lo ideal en ese caso sería eliminar dichas observaciones. Sin embargo, como se está realizando una aplicación de Machine Learning y nuestro principal objetivo es predecir valores, tomaremos la decisión de imputar valor cero a estos NA, teniendo en cuenta que si no reportan experiencia laboral es como si no tuviesen, en nuestro análisis. En el script en R se encuentra el procedimiento.

**1.2.3 Como mínimo, se debe incluir una tabla de estadística descriptiva, pero espero tablas y figuras. Tome esta sección para presentar una narrativa exhaustiva para justificar y defender su elección de datos. Utilice su conocimiento profesional para añadir valor a esta sección. No presente la tabla de manera escueta.**

* 1. ***Perfil edad-ingresos***
     1. **1.3.1 En el set de datos, múltiples variables describen el ingreso. Escoja una que Ud. considere la más representativa para el ingreso total de un trabajador, justificando su selección.**

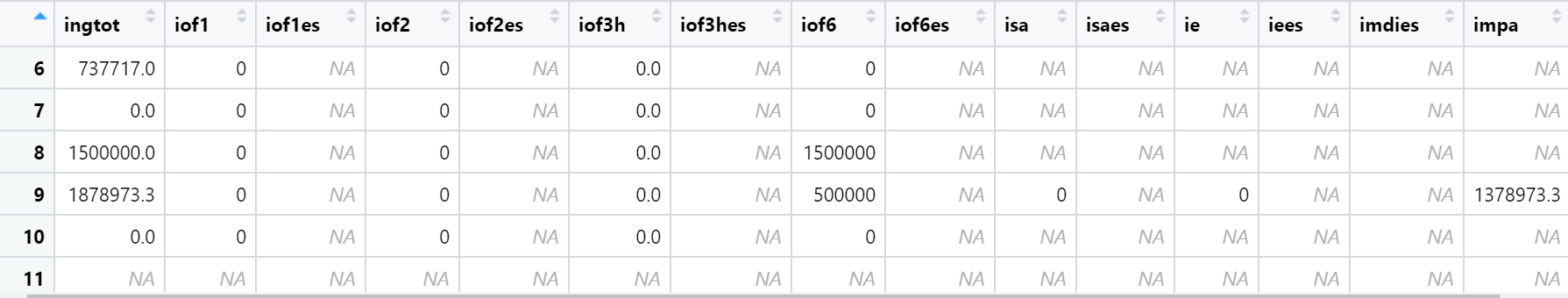
La variable más representativa de los ingresos totales es la denominada “*ingtot”* (Ingresos totales) ya que esta agrupa las diversas variables acerca de los ingresos. Al comparar diferentes variables de los ingresos, se puede observar que los ingresos totales son la suma de cada una del resto de las variables. Por ejemplo:

Se comparó que la Variable *ingtot*  es la suma entre *ingtotes*  e *ingtotob*



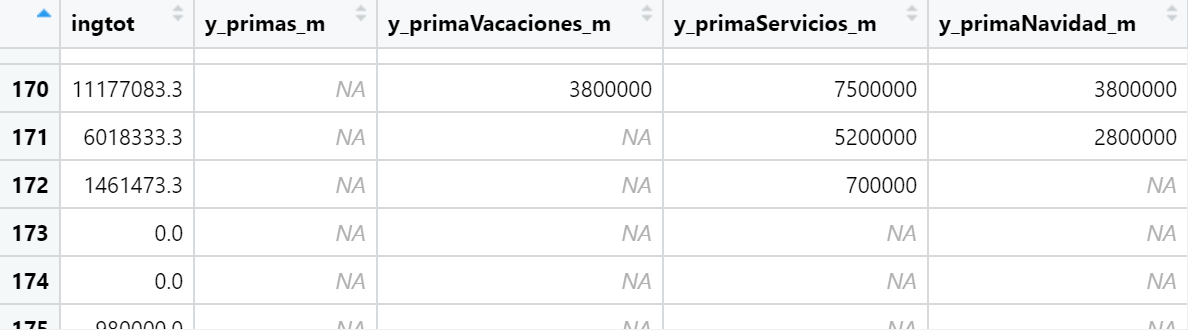
A pesar de tener evidencia de que *ingtot* es la variable representa estas otras dos variables,es pertinente analizar las demás variables que pueden componer el ingreso, siendo estos, los ingresos por intereses, por ayudas, monetarios, arriendos, especie y monetarios para tener certeza acerca de la variable representativa *ingtot.* En la siguiente tabla, se encuentran las variables y su significado

|  |  |
| --- | --- |
| iof1 | Ingreso por intereses y dividendos antes de imputación |
| iof1es | Ingreso por intereses y dividendos imputado (sólo para faltantes o extremos) |
| iof2 | Ingreso por jubilaciones y pensiones antes de imputación |
| iof2es | Ingreso por jubilaciones y pensiones imputado (sólo para faltantes o extremos) |
| iof3h | Ingreso por ayudas de hogares, antes de imputación |
| iof3hes | Ingreso por ayudas de hogares, imputado (sólo para faltantes o extremos) |
| iof3i | Ingreso por ayudas de instituciones, antes de imputación |
| iof3ies | Ingreso por ayudas de instituciones, imputado (sólo para faltantes o extremos) |
| iof6 | Ingreso por arriendos antes de imputación |
| iof6es | Ingreso por arriendos imputado (sólo para faltantes o extremos) |
| isa | Ingreso monetario de la segunda actividad antes de imputación |
| isaes | Ingreso monetario de la segunda actividad imputado (sólo para faltantes o extrem |
| Impa | Ingreso monetario de la primera actividad antes de imputación |
| impaes | Ingreso monetario de la primera actividad imputado (sólo para faltantes, extremo) |

**

Al observar que efectivamente algunas de estas variables de ingreso suman en algunas observaciones el ingreso total se asume que efectivamente la variable que describe todo el ingreso. Sin embargo, para realizar una última comparación, la variable *ingtot* se comparará con las diferentes primas y subsidios:

|  |  |
| --- | --- |
| \_y\_primaServicios\_m | Ingreso prima servicios monetario ultimos 12 meses |
| y\_primaVacaciones\_m | Ingreso prima vacaciones monetario ultimos 12 meses |
| y\_primas\_m | Ingreso primas monetario en el mes |
| y\_salarySec\_m | salary nominal mensual occ. secundario |
| y\_subEducativo\_m | Ingreso subsidio educativo monetario en el mes |
| y\_subFamiliar\_m | Ingreso subsidio familiar monetario en el mes |



Como se puede observar, algunas de estas variables suman en la contabilidad de algunas observaciones del ingreso total. Por lo tanto, tomando como evidencia todas las verificaciones anteriores, se comprueba que la variable que describe el ingreso es *igntot* (Ingreso total), ya que esta contiene todas las demás variables acerca del ingreso que se encuentran en la base de datos de la GEIH

**Nota:** Si desea ver cada una de las tablas sustentadas anteriormente, diríjase al script en R que contiene toda la programación.

* + 1. **Con base en esta estimación utilizando OLS, la ecuación del perfil de edad-ingresos:**

***Ingreso* = β1 + β2Edad + β3Edad^2 + u (2)**

* + 1. **¿Qué tan bueno es este modelo en el ajuste de la muestra?**



La regresión cuenta con la variable *Age2* que representa las edades al cuadrado, ya que se está considerando que después del crecimiento del individuo, llega un punto en el que esa edad genera una relación negativa con el ingreso ya que alcanzó un máximo. Al correr la regresión se presentan estos datos. Los coeficientes representan el impacto que tiene dicha variable en la variable independiente, es decir, ingreso total.

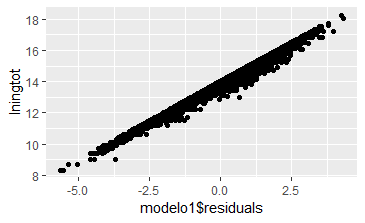
La variable ingtot (Ingreso Total) se transformó para poder analizar correctamente el efecto de los coeficientes de *age* y *age2*  eliminando efecto de dichas unidades y así, lograr la interpretación del modelo porcentual y se los datos sean tratados de manera efectiva y sin inconvenientes en la ejecución de la regresión por MCO.

La constante no suele generar ningún impacto en el modelo, ya que esta es representativa cuando X*i* puede tomar el valor 0. Sin embargo, en este modelo no es posible que las variables tomen ese valor, ya que si fuese el caso, no estarían dentro del modelo. De acuerdo con lo anterior, la constante no genera ningún análisis más allá de ser la intersección que define la relación entre dos variables. (*ingtot* y *age* o *ingtot* y *Age2).*

El coeficiente de *Age* hace referencia a que cuando un individuo aumenta un año de vida, el ingreso de este aumenta en 4.8% su ingreso, contando con un error estándar de la variable de 0.02, siendo este el que mide la precisión con la que cuenta la variable respecto a los valores estimados. A su vez, el coeficiente de *Age2* nos indica una relación negativa entre la variable dependiente y la independiente, es decir por cada año que envejezca el individuo al cuadrado el ingreso disminuye 0.05% , junto con su error estándar de 0.00002, siendo este valor muy pequeño.

Debido a la limpieza de datos generada, se contó con 19801 observaciones y un mismo valor de R y R ajustado de 0.025, representando el poco ajuste que tienen las variables del modelo a la variable independiente, ingresos totales. Por otra parte, se cuenta con el estadístico F con 2 grados de libertad, no rechazando la hipótesis nula de falta de capacidad explicativa de las variables.

Finalmente, cada variable cuenta con una significancia aceptada por el 5%. Esto está representado que se puede tratar a los estimados diferentes de 0. El error estándar residual es de 0.964 siendo este el valor que nos indica que tan bien se están ajustando los datos a la recta de la regresión, aunque este no sea un número muy pequeño, se puede observar que o ajusta todos los datos, pero gran parte de los datos si están cerca de la recta.



* + 1. **Grafique el perfil proyectado de ingresos por edad implícito en la ecuación anterior.**
    2. **¿Cuál es la “edad pico” sugerida por la ecuación anterior? Usar bootstrap para calcular los errores estándar y construir los intervalos de confianza.**
  1. ***La brecha de ganancias***

**Estimar la brecha de ingresos incondicional:**

***log(ingreso)* = β1 + β2Femenino+ u**

* + 1. **¿Cómo debemos interpretar el coeficiente β2? ¿Qué tan bueno es este modelo en samplefit?**

Debido a que la variable “Femenino”, que sería la variable “sex” en nuestra base de datos, se encuentra invertida en la base de datos, es decir, cuando es hombre es igual a 1 y cero cuando es mujer, este β2 es la diferencia en el salario promedio entre hombres y mujeres. Si dejamos que esta variable sea iguala 1 para los hombres como está en la base de datos, significa que, puede ser la diferencia positiva o negativa con respecto a lo que ganan las mujeres.

Para el Samplefit, se observa en la gráfica para la regresión:

* + 1. **Estimar y trazar el perfil de edad-ingresos pronosticado por género. ¿Los hombres y las mujeres en Bogotá tienen la misma intersección y pendientes?**
    2. **¿Cuál es la “edad pico” implícita por género?. Utilice bootstrap para calcular los errores estándar y construir los intervalos de confianza. ¿Se superponen estos intervalos de confianza?**
    3. **Estime una brecha de ingresos condicional que incorpore variables de control tales como características similares del trabajador y del puesto (X).**
       1. **(a) Estime la brecha de ingresos condicional**

**log(*Ingreso*) = β1 + β2Femenino +θX + u**

* + - 1. **Use FWL para repetir la estimación anterior, donde el interés radica en β2. ¿Obtiene las mismas estimaciones?**
      2. **¿Cómo debemos interpretar el coeficiente β2? ¿Qué tan bueno es este modelo en ajuste de muestra? ¿Se reduce la brecha? ¿Es esta evidencia de que la brecha es un problema de selección y no un "problema de discriminación"?**
  1. **Predicción de ganancias.** 
     1. **Divida la muestra en dos muestras: una muestra de entrenamiento (70%) y una muestra de prueba (30%). No olvide establecer una semilla (en R, set.seed(10101), donde 10101 es la semilla).**
        1. **Estime un modelo que solo incluya una constante. Este será el punto de referencia.**
        2. **Estime nuevamente sus modelos anteriores**
        3. **En las secciones anteriores, los modelos estimados tenían diferentes transformaciones de la variable dependiente. En este punto, explora también otras transformaciones de tus variables independientes. Por ejemplo, puede incluir términos polinómicos de ciertos controles o interacciones de estos. Pruebe al menos cinco (5) modelos que aumentan en complejidad.**
        4. **Informe y compare el error de predicción promedio de todos los modelos que estimó anteriormente. Discuta el modelo con el error de predicción promedio más bajo.**
        5. **Para el modelo con el error de predicción promedio más bajo, calcule la estadística de apalancamiento para cada observación en la muestra de prueba. ¿Hay valores atípicos, es decir, observaciones con un alto apalancamiento que impulsen los resultados? ¿Son estos valores atípicos personas potenciales que la DIAN debería investigar, o son simplemente el producto de un modelo defectuoso?**
     2. **Repita el punto anterior pero use la validación cruzada K-fold. Comente las similitudes/diferencias del uso de este enfoque.**
     3. **Con su modelo predicho preferido (el que tiene el promedio más bajo error de predicción) realice el siguiente ejercicio:**

**Escribe un bucle que haga lo siguiente:**

* + - 1. **Estimar el modelo de regresión utilizando todas las observaciones menos la i − ésima.**
      2. **Calcular el error de predicción para la i − ésima observación, es decir, (yi− yˆi)**
      3. **Calcular el promedio de los números obtenidos en el paso anterior para obtener el error cuadrático medio. Esto se conoce como la estadística de validación cruzada Leave-One-Out (LOOCV).**
      4. **Compare los resultados con los obtenidos en el cálculo de la estadística de apalancamiento**